Arabic Text Classification Methods: Systematic Literature Review of Primary Studies

Waleed Alabbas, Haider M. al-Khateeb and Ali Mansour
Institute for Research in Applicable Computing (IRAC)
University of Bedfordshire, Luton, United Kingdom
waleed.alabbas@study.beds.ac.uk, haider.alkhateeb@beds.ac.uk, ali.mansour@beds.ac.uk

***Abstract*— Recent research on Big Data proposed and evaluated a number of advanced techniques to gain meaningful information from the complex and large volume of data available on the World Wide Web. To achieve accurate text analysis, a process is usually initiated with a Text Classification (TC) method. Reviewing the very recent literature in this area shows that most studies are focused on English (and other scripts) while attempts on classifying Arabic texts remain relatively very limited. Hence, we intend to contribute the first Systematic Literature Review (SLR) utilizing a search protocol strictly to summarize key characteristics of the different TC techniques and methods used to classify Arabic text, this work also aims to identify and share a scientific evidence of the gap in current literature to help suggesting areas for further research. Our SLR explicitly investigates empirical evidence as a decision factor to include studies, then conclude which classifier produced more accurate results. Further, our findings identify the lack of standardized corpuses for Arabic text; authors compile their own, and most of the work is focused on Modern Arabic with very little done on Colloquial Arabic despite its wide use in Social Media Networks such as Twitter. In total, 1464 papers were surveyed from which 48 primary studies were included and analyzed.**

Keywords: Arabic text classification; big data; systematic literature review; data mining; Text corpus.

I. INTRODUCTION

Text classification or categorization (TC) is the process of assigning a text document to one or more predefined classes based on their content. TC falls at the crossroad of Machine learning (ML) and information retrieval (IR). There have been tremendous interest in this research area due to the large amount of textual data posted and widely shared on the World Wide Web [1]. TC has been used in different applications including topic identifications, spam filtering and sentiment analysis. While some data can be described as being static such as PDF files, HTML pages could be updated more frequently, while Tweets are retrievable in real-time. In general, Text classifiers can be categorized into two models: Generative and Discriminative. For instance Naïve Bayes (NB) is an example of a generative model that will first try to estimate parameters from p(x | y) and p(y) from the training data, then calculates p(y | x) by using Bayes theorem. Where p(x | y) stands for a conditional probability of x given y is true. It is called "generative" since we can generate new samples by sampling from the learned joint distribution p(x, y). In contrast, a discriminative model estimates parameters of p(x | y) directly from the training data without assuming anything about the input distribution p(x), such models include Support Vector Machines (SVM), Neural Networks and Decision Trees [2, 3]. SVM is considered a non-probabilistic binary linear classifier, it can be used for both classification or regression. For a given set of training samples, the SVM model is representation of these samples as mapped points in space, isolated by a gap to distinguish the different categories. Likewise, Decision Trees can be used as a predictive model. Their structure includes leaves to represent classes (target values) and branches to represent conjunctions of features. However, in complex classification tasks, trees could fail to generalize from the training data (overfitting) or correctly illustrate a concept.

Furthermore, these two approaches can be combined to create a hybrid model, known as Generative-Discriminative Pairs (CDP). It is a relation between a generative model and a discriminative model where one can be directly transformed to the other [3]. Examples include the Discriminative Hidden Markov Model (D-HMM) [4] and the pair of Naive Bayes together with Logistic Regression, in which a model is trained by optimizing a combination of the generative and discriminative log likelihood functions to classify text. CDP can have many advantages to address practical challenges, [5] developed a hybrid model that can switch between generative and discriminative algorithms systematically as a subtask of the learning process, this has allowed them to achieve better results while discovering rare categories in a given dataset.

While discriminative classifiers often outperformance their generative counterparts in accuracy, generative models have several advantages. It is assumed they are easier to classify data and could achieve better accuracy when the training data is limited [3]. However, a generative approach produces a probability density model over all variables in a system and manipulate it to compute classification. While the overall design of generative models has the advantage of being more complete by definition, it can be wasteful and non-robust [6]. A discriminative approach makes no clear attempt to model the underlying distributions of the features in a system and is only interested in optimizing a mapping from the inputs to the required class. As such, learning (not modelling) is the focus of discriminative approaches which often lack flexible modelling, its techniques could feel like black-boxes where the relationships between variables are not as explicit as in generative models [6].

Although TC remains an active research area with novel techniques designed and tested on English scripts [7], there seems to be very little work done on Arabic text. With the absence of a Systematic Literature Review (SLR) based on a

W. Alabbas, H. M. Al-Khateeb and A. Mansour, "Arabic text classification methods: Systematic literature review of primary studies," 2016 4th IEEE International Colloquium on Information Science and Technology (CiSt), Tangier, 2016, pp. 361-367. doi: 10.1109/CIST.2016.7805072

comprehensive search protocol and quality assessment, it is not possible to determine the research gap for Arabic text, this has become one of the objectives for this study. For instance, it is important to conclude better performing classifiers and which text pre-processing and Dimensionality Reduction Techniques (DRT) [8] were proven more effective for Arabic.

Arabic is the 5th widely used language in the world. It is officially used in 24 countries, the mother tongue for more than 422 million persons and the second language for almost another 250 million. Arabic has 28 letters and the orientation of writing is from right to lift. Its script has a unique shape, marks, diacritics, Style (font), numerals, distinctive letters and none distinctive letters [9]. Noaman and Al-Ghuribi, [10] discussed its complex morphology and how words could have different meaning within a given context. Arabic is highly inflectional and derivational [11], it does not use capitalization for proper nouns which is a very useful input when classifying English documents. Arabic synonyms are widespread [12]. The majority of words have a tri-letter root, while the rest have a quad-letter root, penta-letter root or hexa-letter root [13].

Recent publications into this growing area of research include Fawaz and AbuZein's work [14] to enhance classifier's performance on Arabic text using cosine similarity and latent semantic indexing, the effect of preprocessing on Arabic document categorization by Ayedh et al.[15] and others [16-18]

The remaining of this paper covers the methodology in Second 2 which also discusses the research questions of this study, protocol used and finally the data extraction strategy. Section 3 contains SLR results analysis and discussion of key findings from the included primary studies. Finally, conclusions are written in Section IV.

II. METHODOLOGY

The research method is based on the SLR guidelines for the discipline of computer engineering as proposed by Kitchenham and Charter [19]. Key phases we followed are demonstrated in Fig 1, we share further reflection on each within the consequent sections. In general, we have identified the problem statement, research questions and fundamental aspects of the review protocol as part of the Planning phase. To mitigate subjectivity, we enforced a role that each of these phases is initiated after a full evaluation and approval of the previous one. The Search Strategy, consisted of the study selection criteria, procedure, unified search string and study quality assessment. The third phase is mainly conserved with the development of our Data Extraction strategy. And the final phase of the systematic review involved data synthesis and critical analysis.

Figure 1 – Main stages followed in this SLR.

A. Research questions

The main aim of this study can be achieved through answering the research questions define and discussed below:

RQ1. What TC models have been applied on Arabic text and supported by an empirical evidence to estimate their accuracy? And which models performed better on Arabic text?

RQ2. What characteristics can be identified to describe corpuses, techniques and algorithms used that can affect accuracy for these TC models?

The term 'models' used in the questions above could be used interchangeably with 'methods' and 'techniques'. Answering RQ1 helps to conclude a list of all relevant TC methods within the scope and requirement of this study, while RQ2 investigates their key characteristics. RQ1 helps to research the accuracy of their implementation and therefore reliability in a real life application. Both RQ1 and RQ2 help to identify the gap in current literature and suggest areas for further investigation.

To frame these research questions effectively, PICOC criteria (Population, Intervention, Comparison, Outcome, and Context) [19, 20] were applied from viewpoint of software engineering as follows:

Population	Text Classification Models.
Intervention	Generative and Hybrid models.
Comparison	Discriminative models.
Outcomes	Accuracy of the models analysed.
Context	Academic research.

B. Data sources and search strategy

Pioneer database sources for software engineering research publications have been used as shown in Table 1. This study begun in January 2015 and therefore considered publications up to that date. Searching keywords were defined to include the following key terms and synonyms constructed with logical operators to return the best possible search outcome:

('Arabic text' OR 'Arabic script') AND ('classification' OR 'Classifier' OR 'categorization' OR 'categorisation').

This search string was adapted to the built-in options of each database from Table 1 to filter and refine results. Further, grey literature was considered in our search strategy together with a snow balling approach (reference of references) where any paper collected by our search criteria can manually lead to another reference from within its bibliography.

TABLE 1 – DATABASES

Database	URL
IEEExplore	http://ieeexplore.ieee.org
ACM Digital library	http://dl.acm.org
CiteSeerX library	http://citeseerx.ist.psu.edu/index
Science Direct	http://www.sciencedirect.com
Springer	http://link.springer.com/
Academic Search Elite	https://www.ebscohost.com/
DOAJ	https://doaj.org/
Web of Knowledge	http://www.webofknowledge.com
Scopus	http://www.scopus.com/
Google scholar	http://scholar.google.co.uk

W. Alabbas, H. M. Al-Khateeb and A. Mansour, "Arabic text classification methods: Systematic literature review of primary studies," 2016 4th IEEE International Colloquium on Information Science and Technology (CiSt), Tangier, 2016, pp. 361-367. doi: 10.1109/CIST.2016.7805072

C.Study selection criteria

In this step we apply rigorous inclusion and exclusion criteria to ensure valuable and relevant information in response to our defined research questions. These criteria were enforced after reading the title, abstract and then full text of the articles as demonstrated in the study selection procedure shown in (2.4.) For instance, [21] was excluded because it does not report the method's accuracy and [22] was not a primary study.

Inclusion criteria:

- Must be a primary study reporting on TC models from the area of software engineering /data mining.
- Must address the accuracy of the TC model/method.
- Must include analysis and empirical evidence.

Exclusion criteria:

- Publication is not peer reviewed.
- Arabic is not the language used to test the TC model.

D.Study selection procedure

The selection of the primary studies was examined by all authors. Four different phases show how the selection procedure was implemented as illustrated in Fig 2:

Phase 0 – Keywords-based filtering.

In this phase, the search string was applied to the ten scholarly databases shown in Table 1. This has yielded a total of 1464 articles which were included in the next phase.

Phase 1 –Title, indexing keywords and abstract-based filtering.

In this phase, titles were examined against the inclusion and exclusion criteria. Articles deemed to be of any relevance were directly included in the next phase. In conclusion, 863 articles were discarded and 365 articles were included.

Phase 2 – Full text-based filtering.

This was the final stage where the reviewers discussed and resolved disagreements regarding the relevance of the articles to the study. A total of 192 articles were identified to be duplicates download from different databases and were therefore discarded. Upon reconsideration of the inclusion and exclusion criteria, 125 articles were excluded for different reasons; for instance, [23] was not peer reviewed, [24] did not include an empirical study and [25] did not satisfy a number of the quality assessment criteria shown in (2.5). The final set of primary study had a total of 48 remaining articles.

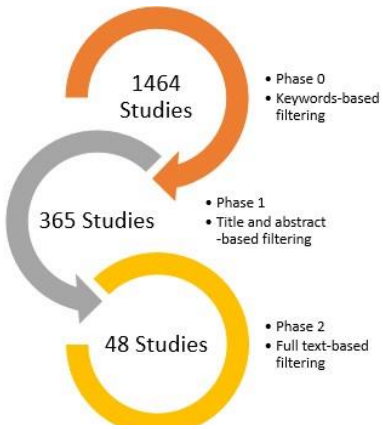

Figure 2 – The number of primary studies included in each phase of the study selection procedures.

E.Study quality assessment

Included papers had to satisfy a quality assessment designed as a measure to determine if a given paper is suitable to address our research questions. The following checklist had to be met with affirmative answers:

- Was the number of training and testing data identified?
- Were the pre-processing techniques used in the study clearly described and their selection justified?
- Were the classifiers used in study clearly described?
- Is there comparison with other approaches?
- Were the performance measures fully defined?

F.Data extraction strategy

Data extracted from the studies, were tabulated and comprised the following characteristics: year of publication, number of learning and testing documents, features selection approaches, classification algorithm and accuracy.

III. Results Analysis and Discussion

A. *Primary studies*

There were a total number of 48 included studies in the form of journal articles and conference proceedings published between 2006 and 2014. Our analysis shows most articles were published within the last 5 years which was an early indication that Arabic TC is an active research area and started to evolve very recently. More details are demonstrated in Table 2.

TABLE 2 – THE DISTRIBUTION OF PRIMARY STUDIES BY PUBLICATION TYPE [JOURNAL (J) OR CONFERENCE (C)] AND PUBLICATION YEAR

Source	06	07	08	09	10	11	12	13	14	Total
J	1	2	2	3	2	4	3	1	7	25
C	1	0	2	2	5	3	4	6	0	23
%	4	4	7	10	15	15	15	15	15	100

B. *Key focus areas*

We found that primary studies can be classified by their main focus area into four domains: TC algorithms, Features Selection (FS), Stemming Techniques (ST) and Term Weighting (TW). The majority of work was on TC algorithms as shown in Table 3.

TABLE 3 – KEY FOCUS AREA FOR INCLUDED PAPERS

Focus area	%	Studies
TC algorithms	56	[11, 26-51]
FS	25	[12, 52-62]
ST	13	[63-68]
TW	6	[13, 69, 70]

Each of these focus areas are discussed in details within Sections III.D, III.E and III.F.

W. Alabbas, H. M. Al-Khateeb and A. Mansour, "Arabic text classification methods: Systematic literature review of primary studies," 2016 4th IEEE International Colloquium on Information Science and Technology (CiSt), Tangier, 2016, pp. 361-367. doi: 10.1109/CIST.2016.7805072

C. Data collection (Corpus)

Collecting data to create a suitable dataset is the first step in text classification studies. Whilst there are several free benchmarking datasets for English used for TC purposes: the 20 Newsgroup contains around 20,000 texts distributed almost evenly into 20 classes; Reuters-21578 contains 21,578 texts belonging to 17 classes; and RCV1 (Reuters Corpus Volume 1), contains 806,791 texts classified into four main classes [26]. Unfortunately, the case is different for Arabic. There seems to be no free benchmarking dataset identified from the included studies for Arabic TC. For most research, authors collect data to build their very own datasets, mostly from online formal websites and news articles. Table 4 describes the datasets used in each study. It also shows the language model selected, whether it is classical Arabic (also known as Quaranic) which could also include old poem and religious scripts; modern Arabic currently used in formal press and government communications; colloquial Arabic as in informal local dialects; or a mixture of these. It has also been noted that some papers do not seem to describe their datasets enough which makes it difficult to classify their datasets. Such works usually do not publish their data for other researchers to utilize. Consequently, the confidence in the results derived from such experimental studies is not satisfactory enough. The performance of the adopted data mining approaches is biased to such data sets and could be ambiguous.

TABLE 4 – SOURCES FOR BUILDING DATASETS

Models	Corpus	Studies
Classic	Quran	[49]
	Religious scripts	[38, 45, 55, 57]
	Old books	[31]
Modern	Websites	[13, 27, 29, 53, 60, 61, 63, 64, 66, 67]
	News articles	[11, 12, 28, 32, 39, 42-44, 46, 47, 50-52, 54, 59, 65, 68-70]
Colloquial	User Reviews	[48]
Hybrid		[26, 30, 34, 41]
unknown		[33, 35-37, 40, 56, 58, 62]

Results show that most work is conducted on the modern language with a single study [48] covering informal (colloquial) writing, this is an interesting finding because it recovers a huge and critical technology gap, informal Arabic is people use on social media, especially Twitter. Arabic dialects vary from one Arab country to another and could also slightly vary between cities and towns.

With regards to the size of datasets, they ranged from 119 documents divided into three classes [12] to 17,652 documents divided into six classes [30]. The vast majority of studies measures the size by the number of documents rather than word count. This detail given an indication on the size but it remains a challenge to have an accurate statistical comparison between the different datasets used.

D. Text pre-processing and dimensionality reduction techniques

Pre-processing is a trial to improve text classification by removing worthless data. It may include the removal of numbers, punctuation (e.g. hyphens) and stop-words (e.g. prepositions and pronouns). Due to its writing style, Arabic requires careful strategies at this stage to normalize writing forms and removing diacritics.

A number of dimensionality reduction techniques are also used to reduce the number of terms included for analysis (classification); high dimensionality data do not satisfy the requirements of TC methods to produce reasonably accurate outcome and are therefore considered problematic [71]. Included studies identified the use of two reduction techniques, namely: Stemming and Feature Selection.

1) Stemming

Stemming is a technique to reduce the high dimensionality of the feature space in text classification. Several Stemming approaches exist for the Arabic language each produces a different set of roots. These are identified in Table 5 and discussed in further details below.

Root-based stemming (Lexical) is based on removing all attached prefixes and suffixes in an attempt to extract the root of a given Arabic surface word. An example of this approach is the Khoja stemmer [72]. Its core-function works by mapping words into their root patterns. Root patterns in Arabic are three, four, five, or six-letter patterns. More than 80% of the Arabic words can be mapped into three-letter root pattern, reducing a word to its root pattern could decreases the number of words from hundreds of thousands to as little as 4,749 as in [69].

TABLE 5 – STEMMER TECHNIQUES USED

Stemmer	Studies
Root-based stemming	[12, 34, 35, 39, 48, 65, 69]
Light stemming	[27, 29, 43, 45, 47, 55, 60, 64, 66]
Statistical stemmer	[49, 61, 62]
Hybrid	[63]

Light Stemming does not attempt to give the linguistic root pattern for the word, instead, its main focus is to remove the most frequent suffixes and prefixes. There are different types of Light Stemming and many studies have considered this approach (Table 5). The literature in general gives an argument that light stemming allows remarkably good information retrieval, [73] discuss this in further details.

Statistical stemmer (character level N-Gram), N-Gram is a set of N consecutive characters extracted from a word. The main idea behind this approach is that, similar words will have a high proportion of N-Gram in common. Typical values for n are 2 or 3, these corresponding to the use of diagrams or trigrams, respectively. For example, when 3-grams is applied on the following string: “text classification”, the output is: “tex”, “ext”, “xt_”, “t_c”, “_cl “, “ cla”, “las”, “ass”, and so on [63]. Each of these strings will then be compared against the output of another string to measure and determine the level of similarity between the two.

A hybrid approach was also tested where a number of stemming techniques are used together in an attempt to improve the process. For example [63] proposed a hybrid method incorporating Khoja stemmer, light stemmer and N-Gram. Results were promising with an improvement in the overall accuracy. Likewise [69] used root extraction by assigning weights and ranks to the letters that constitute a

word. However, they mention that roots are semantically weak in the meaning that several words can be mapped onto the same root.

Nonetheless, in some cases stemming techniques could decrease the performance of the classifier used. Kanaan et al., [44] observed this behavior when light stemming was used with the Rocchio and NB algorithms. Likewise, Al-Kabi et al., [67] conducted an experiment and concluded that Khoja stemmer did not improve the classification accuracy for NB, SVM (SOM) and decision tree (J48).

2) Feature selection

Some reduction methods utilize features (terms) selection to reduce dimensionality. These statistical techniques work at the term level, as such, when 3-gram is utilized; text is split into chunks of 3 terms (words rather than characters). Table 6 demonstrates which FS techniques was used by each study.

TABLE 6 – FEATURE SELECTION TECHNIQUES

FS Techniques	Studies
Chi-square	[29, 31, 41, 47, 52, 54, 56, 58, 61]
Term Frequency	[43, 59]
Document Frequency	[70]
Information Gain	[30]
N-gram	[13, 48, 68]
Hybrid	[26]

Most studies applied Chi-square (CHI) while there was a single study [26] attempting a hybrid approach in which the authors applied Document Frequency and Galavotti, Sebastiani, Simi (GSS).

E. Feature reresentation (term weighting)

TC algorithms require that text features are formatted before they can be interpreted by the specified classifier, this process is also referred to as term weighting because each term is entered together with a weight value. Included papers show the most used technique is the Term Frequency-Inverse Document Frequency (TF-IDF) as in [27, 32, 37, 40, 43, 45, 48, 51, 53, 55, 57, 58, 60-62, 67]. It is a statistical method to indicate the significance of a word within a given corpus. This utilization of the technique is justified assuming the authors wanted to weight terms while considering its significance across all documents rather than a single one. Although, in [58] a simpler but more limited method has also been used to conclude a Boolean value of zero or one, a term can be described to be either important or not important. Whilst in TF-IDF, for a given term, a bigger TF-IDF value indicates a more frequent word. As such, data can be represented as a matrix with *n* rows and *m* columns wherein the rows correspond to the texts in the training data, and the columns correspond to the selected feature. The value of each cell in this matrix represents the weight of the feature in the text.

F. Classification algorithms and accuracy

Each study used their very own corpus and different experiment conditions in terms of their training and testing procedure, pre-processing and DRT. Hence, it is not feasible to statistically compare accuracy values (cross studies). However, when we analyze the outcome of different studies, there is evidence that the Support Vector Machine (SVM) classifier (a discriminative model) outperforms other classifiers with the exception of two studies reporting in favor of the C5.0 Decision Tree Algorithm, and one study on k-NN. This outcome is demonstrated in Table 7.

TABLE 7 – STUDIES INVESTIGATING ACCURACY. ACCURACY VALUES FOR EACH STUDY HAVE BEEN REPORTED IN THE FOLLOWING FORMAT: [STUDY] (ACCURACY FOR THE PREEMINENT CLASSIFIER – ACCURACY FOR THE FIRST METHOD IN COMPARISION, ACCURACY FOR THE SECOND METHOD, …)

Preeminent Classifier	Compared with	Studies (and accuracy values)
SVM	NB	[26](0.805-0.755), [36](0.778-0.74), [38](0.954-0.884), [42](0.778-0.74), [61](0.9241-0.8949), [65] (0.8638-0.7741)
	k-NN	[46](0.827-0.448),
	ANN	[27] (0.956- 0.94)
	NB, k-NN, ROCHIO	[56] (0.9141-0.8778, 0.7581, 0.7472)
	J48, NB	[33](0.948-0.8942, 0.8507), [37](0.9608-0.9048, 0.856), [67](0.896-0.753, 0.835)
	J48, NB, k-NN	[47] (0.98-0.856, 0.967, 0.799)
	NB, k-NN	[48] (0.611-0.585, 0.601), [51](0.914-0.845, 0.727)
NB	ANN, k-NN	[28] (0.85-0.81)
	k-NN	[39] (0.81-0.78), [58](0.8574-0.7995)
	k-NN, RACHIO	[44] (0.82-0.7871, 0.7882)
	SVM, k-NN	[48] (0.857-0.824, 0.646)
Decision-tree (C5.0)	SVM, NB, ANN	[30] (0.8443-0.761, 0.7566, 0.6378)
	SVM	[41] (0.7842-0.6865)
k-NN	SVM, NB	[48] (0.666-0.598, 0.563)

While all included studies have also reported the accuracy of their classifiers, Table 7 includes only those attempted to conduct experiments on multiple algorithms within a controlled environment for comparison purposes.

Results show that generative models remain an option when the amount of training is relatively small and could therefore be faster, both algorithms that reportedly outperformed other models are discriminative (SVM and C5.0). SVM is a supervised learning algorithm, with an appropriate kernel, the algorithm can function competently whether or not the data is linearly separable. It is widely used even with text of high-dimensionality. However, its disadvantage can be summarized to be the algorithms complexity, interpretability and memory requirements [74]. However, not all discriminative models performed better, the K-Nearest Neighbor (k-NN) Classifier is an exemplar for this case. It is discriminative because it models the conditional probability of data belonging to a given class. k-NN computes the similarities between a new sample and the training samples previously stored in a dataset. The most K similar ones are then listed in a descending order. Finally, the new sample takes the class label that belongs to the majority of these K neighbors [43]. It should therefore not be preferred for text categorization [74]. Nonetheless, although C5.0 Decision Tree algorithm outperformed SVM, the later outperformed another Decision

Tree algorithm; J48 while many other remain untested in the literature.

As mentioned earlier, the remaining set of the included studies did not conduct a comparison between classifiers, they have instead investigated other factors. For instance, [63] used NB with different stemmer techniques and found that a hybrid method gives more accurate results if compared to a root-based stemmer, light stemmer or n-gram (statistical stemming). Likewise, [47] used SVM with different stemming techniques, however the study reports very minor effect on accuracy.

IV. CONCLUSION

More work need to be done on Arabic text analysis as it can be applied to solve real-world problems such as automating procedures, building intelligence and mitigating cybercrime [75]. Future work on TC techniques for Arabic text should ideally consider using a corpus that is available online for download; this will enable comparative experiments by other researchers and conclude robust facts with regards to the accuracy and speed of the different algorithms and techniques available. Additionally, datasets should be described thoroughly in the papers, sharing the word count to describe the size is the right approach rather than the number of documents collected.

Implementing a hybrid Stemming and/or Feature Selection approach could improve the accuracy as few studies suggest. Majority of papers report on using root-based stemming, light stemming and Chi-square, therefore more research is needed to investigate the opportunities and threats for adopting hybrid Dimensionality Reduction Techniques on Arabic text during both: Stemming and Feature Selection.

Not all discriminative algorithms outperform the accuracy of generative models; NB outperformed k-NN, however both preeminent algorithms from the included studies were discriminative; SVM and C5.0. Additionally, no work has be found by our search protocol to compare with a hybrid model; Generative-Discriminative Pairs (CDP).

Further, TF-IDF was used in the vast majority of papers but there was little discussion and justification for adopting this statistical method, it is very critical that new research realize this limitation in current literature, lack of details was a key reason to exclude some papers in our protocol mainly because they have failed to describe their training datasets and report the accuracy of the utilized algorithms.

REFERENCES

[1] Chakravarty, S.: 'A survey on text classification techniques for e-mail filtering'. Proc. Machine Learning and Computing (ICMLC), 2010 Second International Conference on 2010

[2] Jordan, A.: 'On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes', Advances in neural information processing systems, 2002, 14,

[3] Raina, R., Shen, Y., McCallum, A., and Ng, A.Y.: 'Classification with hybrid generative/discriminative models'. Proc. Advances in neural information processing systems 2003

[4] Xue, J.: 'Aspects of generative and discriminative classifiers', University of Glasgow, 2008

[5] Hospedales, T.M., Gong, S., and Xiang, T.: 'Finding rare classes: Active learning with generative and discriminative models', Knowledge and Data Engineering, IEEE Transactions on, 2013, 25, (2),

[6] Jebara, T.: 'Discriminative, generative and imitative learning', 2001,

[7] Singla, A., Patra, S., and Bruzzone, L.: 'A novel classification technique based on progressive transductive SVM learning', Pattern Recognition Letters, 2014, 42,

[8] Tu, S., and Xu, L.: 'A theoretical investigation of several model selection criteria for dimensionality reduction', Pattern Recognition Letters, 2012, 33, (9),

[9] Odeh, A., Abu-Errub, A., Shambour, Q., and Turab, N.: 'Arabic Text Categorization Algorithm using Vector Evaluation Method', arXiv preprint arXiv:1501.01318, 2015,

[10] Noaman, A., and Al-Ghuribi, S.: 'A NEW APPROACH FOR ARABIC TEXT CLASSIFICATION USING LIGHT STEMMER AND PROBABILITIES', International Journal of Academic Research, 2012, 4, (3),

[11] El-Halees, A.M.: 'Arabic text classification using maximum entropy', The Islamic University Journal (Series of Natural Studies andEngineering) 2007, Vol. 15,No.1,pp 157-167, 2007, ISSN 1726-6807, http//www.iugzaza.edu.ps/ara/research/,

[12] Saad, M.K., and Ashour, W.: 'Arabic text classification using decision trees'. Proc. Proceedings of the 12th international workshop on computer science and information technologies CSIT 2010

[13] Khreisat, L.: 'A machine learning approach for Arabic text classification using N-gram frequency statistics', Journal of Informetrics, 2009, 3, (1),

[14] Al-Anzi, F.S., and AbuZeina, D.: 'Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing', Journal of King Saud University-Computer and Information Sciences, 2016,

[15] Ayedh, A., TAN, G., Alwesabi, K., and Rajeh, H.: 'The Effect of Preprocessing on Arabic Document Categorization', Algorithms, 2016, 9, (2),

[16] Kanan, T., and Fox, E.A.: 'Automated Arabic Text. Classification with P-Stemmer', Machine Learning, and a Tailored News Article Taxonomy. J. Assoc. Inf. Sci. Technol, 2016,

[17] Mohammad, A.H., Al-Momani, O., and Alwada'n, T.: 'Arabic Text Categorization using k-nearest neighbour, Decision Trees (C4. 5) and Rocchio Classifier: A Comparative Study', 2016,

[18] Abainia, K., Ouamour, S., and Sayoud, H.: 'Topic identification of Arabic noisy texts based on KNN'. Proc. Information and Communication Technology Research (ICTRC), 2015 International Conference on 2015

[19] Kitchenham, B., and Charters, S.: 'Guidelines for performing systematic literature reviews in software engineering': 'Technical report, Ver. 2.3 EBSE Technical Report. EBSE' (2007),

[20] Higgins, J.P.T., and Green, S.: 'Cochrane handbook for systematic reviews of interventions' (Wiley Online Library, 2008. 2008)

[21] Al-Shawakfa, E., Al-Badarneh, A., Shatnawi, S., Al-Rabab'ah, K., and Bani-Ismail, B.: 'A comparison study of some Arabic root finding algorithms', Journal of the American Society for Information Science and Technology, 2010, 61, (5),

[22] Mamoun, R., and Ahmed, M.A.: 'A Comparative Study on Different Types of Approaches to the Arabic text classification', 1st International Conference of Recent Trends in Information and Communication Technologies, 2014,

[23] Alaa, E.: 'A Comparative Study on Arabic Text Classification', researchgate.net, 2008,

[24] Yahia, M.E.: 'Arabic text categorization based on rough set classification'. Proc. Computer Systems and Applications (AICCSA), 2011 9th IEEE/ACS International Conference on 2011

[25] Ben Othmane Zribi, C., Ben Fraj, F., and Ben Ahmed, M.: 'Combining classifiers for supertagging Arabic texts'. Proc. Natural Language Processing and Knowledge Engineering (NLP-KE), 2010 International Conference on 2010

[26] Khorsheed, M.S., and Al-Thubaity, A.O.: 'Comparative evaluation of text classification techniques using a large diverse Arabic dataset', Language Resources and Evaluation, 2013, 47, (2),

[27] Belkebir, R., and Guessoum, A.: 'A Hybrid BSO-Chi2-SVM Approach to Arabic Text Categorization', 2013 Acs International Conference on Computer Systems and Applications (Aiccsa), 2013,

[28] AlSaleem, S.M.: 'Neural Networks for the Automation of Arabic Text Categorization', 2013 International Conference on Computer Applications Technology (Iccat), 2013,

[29] Kadhim, M.H., and Omar, N.: 'Automatic Arabic Text Categorization using Bayesian Learning', 2012 7th International Conference on Computing and Convergence Technology (Iccct2012), 2012,

W. Alabbas, H. M. Al-Khateeb and A. Mansour, "Arabic text classification methods: Systematic literature review of primary studies," 2016 4th IEEE International Colloquium on Information Science and Technology (CiSt), Tangier, 2016, pp. 361-367. doi: 10.1109/CIST.2016.7805072

[30] Al-Thubaity, A., Almuhareb, A., Al-Harbi, S., Al-Rajeh, A., and Khorsheed, M.: 'KACST Arabic Text Classification Project: Overview and preliminary results' (2008. 2008)
[31] Altheneyan, A.S., and Menai, M.E.B.: 'Naïve Bayes classifiers for authorship attribution of Arabic texts', Special Issue on Arabic NLP, 2014, 26, (4),
[32] Hmeidi, I., Hawashin, B., and El-Qawasmeh, E.: 'Performance of KNN and SVM classifiers on full word Arabic articles', Intelligent computing in engineering and architecture, 2008, 22, (1),
[33] Majed, I.H., Fekry, O., Minwer, A.L.d., and Shamsan, A.: 'Arabic Text Classification using Smo,naĬve Bayesian, J48 Algorithms', International Journal of Research and Reviews in Applied Sciences, 2011, (2),
[34] Thabtah, F., Gharaibeh, O., and Abdeljaber, H.: 'Comparison of rule based classification techniques for the Arabic textual data'. Proc. Innovation in Information & Communication Technology (ISIICT), 2011 Fourth International Symposium on 2011
[35] Al-Jaloud, F., Hezam, R.B., and Aoun-Allah, M.: 'Classifying Arabic web pages toolkit'. Proc. Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics 2012
[36] Alwedyan, J., Hadi, W.e.M., Salam, M.a., and Mansour, H.Y.: 'Categorize arabic data sets using multi-class classification based on association rule approach'. Proc. Proceedings of the 2011 International Conference on Intelligent Semantic Web-Services and Applications 2011
[37] Al-Shargabi, B., Al-Romimah, W., and Olayah, F.: 'A comparative study for Arabic text classification algorithms based on stop words elimination'. Proc. Proceedings of the 2011 International Conference on Intelligent Semantic Web-Services and Applications 2011
[38] Hadi, W.e.M., Salam, M.a., and Al-Widian, J.A.: 'Performance of nb and svm classifiers in islamic arabic data'. Proc. Proceedings of the 1st International Conference on Intelligent Semantic Web-Services and Applications 2010
[39] Duwairi, R.M.: 'Arabic Text Categorization', The International Arab Journal of Information technology, 2007, 4, (2),
[40] Abbas, M., Smaili, K., and Berkani, D.: 'Tr-classifier and knn evaluation for topic identification tasks', The International Journal on Information and Communication Technologies (IJICT), 2010, 3, (3),
[41] Al-Harbi, S., Almuhareb, A., Al-Thubaity, A., Khorsheed, M.S., and Al-Rajeh, A.: 'Automatic Arabic text classification', In, Proceedings of The 9th International Conference on the Statistical Analysis of Textual Data, Lyon-, France., 2008,
[42] Alsaleem, S.: 'Automated Arabic Text Categorization Using SVM and NB', Int.Arab J.e-Technol., 2011, 2, (2),
[43] Al-Shalabi, R., Kanaan, G., and Gharaibeh, M.: 'Arabic text categorization using kNN algorithm'. Proc. Proceedings of The 4th International Multiconference on Computer Science and Information Technology 2006
[44] Kanaan, G., Al-Shalabi, R., Ghwanmeh, S., and Al-Ma'adeed, H.: 'A comparison of text-classification techniques applied to Arabic text', Journal of the American Society for Information Science and Technology, 2009, 60, (9),
[45] Harrag, F., and Al-Qawasmah, E.: 'Improving Arabic Text Categorization Using Neural Network with SVD', JDIM, 2010, 8, (4),
[46] Al-Thwaib, E.: 'Support Vector Machine versus k-Nearest Neighbor for Arabic Text Classification', International Journal of Sciences, 2014, 3, (2014-06),
[47] Hmeidi, I., Al-Ayyoub, M., Abdulla, N.A., Almodawar, A.A., Abooraig, R., and Mahyoub, N.A.: 'Automatic Arabic text categorization: A comprehensive comparative study', Journal of Information Science, 2014,
[48] Duwairi, R., and El-Orfali, M.: 'A study of the effects of preprocessing strategies on sentiment analysis for Arabic text', Journal of Information Science, 2014, 40, (4),
[49] Mahafdah, R., Omar, N., and Al-Omari, O.: 'Arabic Part of Speech Tagging Using K-Nearest Neighbour and Naive Bayes Combination ', Journal of Computer Science, 2014, 10, (11),
[50] Al-diabat, M.: 'Arabic text categorization using classification rule mining', Applied Mathematical Sciences, 2012, 6, (81),
[51] Zrigui, M., Ayadi, R., Mars, M., and Maraoui, M.: 'Arabic text classification framework based on latent dirichlet allocation', CIT.Journal of Computing and Information Technology, 2012, 20, (2),
[52] Al-Thubaity, A., Abanumay, N., Al-Jerayyed, S., Alrukban, A., and Mannaa, Z.: 'The Effect of Combining Different Feature Selection Methods on Arabic Text Classification', 2013 14th Acis International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/distributed Computing (Snpd 2013), 2013,
[53] Al Zaghoul, F., and Al-Dhaheri, S.: 'Arabic Text Classification Based on Features Reduction Using Artificial Neural Networks' (Computer Modelling and Simulation (UKSim), 2013 UKSim 15th International Conference on. IEEE, 2013., 2013. 2013)
[54] Elberrichi, Z., and Abidi, K.: 'Arabic Text Categorization: a Comparative Study of Different Representation Modes', International Arab Journal of Information Technology, 2012, 9, (5),
[55] Harrag, F., and El-Qawasmah, E.: 'Neural Network for Arabic Text Classification' (In Applications of Digital Information and Web Technologies, 2009. ICADIWT'09. Second International Conference on the (pp. 778-783). IEEE., 2009. 2009)
[56] Mesleh, A.M.d.: 'Feature sub-set selection metrics for Arabic text classification', Pattern Recognition Letters, 2011, 32, (14),
[57] Harrag, F., El-Qawasmah, E., and Al-Salman, A.M.S.: 'Comparing Dimension Reduction Techniques for Arabic Text Classification Using BPNN Algorithm'. Proc. Integrated Intelligent Computing (ICIIC), 2010 First International Conference on 2010
[58] Al-Thubaity, A., Alanazi, A., Hazzaa, I., and Al-Tuwaijri, H.: 'Weirdness Coefficient as a Feature Selection Method for Arabic Special Domain Text Classification'. Proc. Asian Language Processing (IALP), 2012 International Conference on 2012
[59] Al-Thwaib, E.: 'Text summarization as feature selection for arabic text classification', World of Computer Science and Information Technology Journal (WCSIT), 2014, 4, (7),
[60] Duwairi, R.M.: 'Statistical Feature Selection Techniques for Arabic Text Categorization'. Proc. The Fourth International Conference on Information and Communication Systems (ICICS 2013), , Irbid, Jordan, April, 23-25, 2013
[61] Raheel, S., and Dichy, J.: 'An Empirical Study on the Feature's Type Effect on the Automatic Classification of Arabic Documents': 'Computational Linguistics and Intelligent Text Processing' (Springer, 2010),
[62] Al-Shalabi, R., and Obeidat, R.: 'Improving KNN Arabic text classification with n-grams based document indexing'. Proc. Proceedings of the Sixth International Conference on Informatics and Systems, Cairo, Egypt, 2008
[63] Hadni, M., Lachkar, A., and Alaoui Ouatik, S.: 'A New and Efficient Stemming Technique for Arabic Text Categorization', 2012 International Conference on Multimedia Computing and Systems (Icmcs), 2012,
[64] Duwairi, R., Al-Refai, M.N., and Khasawneh, N.: 'Feature Reduction Techniques for Arabic Text Categorization', Journal of the American Society for Information Science and Technology, 2009, 60, (11),
[65] Al-Shammari, E.T.: 'Improving Arabic document categorization: Introducing local stem'. Proc. Intelligent Systems Design and Applications (ISDA), 2010 10th International Conference on 2010
[66] Omer, M.A.H., and Ma, S.-L.: 'Stemming algorithm to classify Arabic documents', Journal of Communication and Computer, 2010, 7, (9),
[67] Al-Kabi, M., Al-Shawakfa, E., and Alsmadi, I.: 'The Effect of Stemming on Arabic Text Classification: An Empirical Study', Information Retrieval Methods for Multidisciplinary Applications, 2013,
[68] Nehar, A., Ziadi, D., and Cherroun, H.: 'Rational kernels for arabic text classification': 'Statistical Language and Speech Processing' (Springer, 2013),
[69] Duwairi, R.M.: 'Machine learning for Arabic text categorization', Journal of the American Society for Information Science and Technology, 2006, 57, (8),
[70] Ababneh, J., Almomani, O., Hadi, W., El-Omari, N.K.T., and Al-Ibrahim, A.: 'Vector Space Models to Classify Arabic Text', International Journal of Computer Trends and Technology (IJCTT), 2014, 7, (4),
[71] Sebastiani, F.: 'Machine learning in automated text categorization', ACM computing surveys (CSUR), 2002, 34, (1),
[72] Khoja, S.: 'APT: Arabic part-of-speech tagger'. Proc. Proceedings of the Student Workshop at NAACL 2001
[73] Larkey, L.S., Ballesteros, L., and Connell, M.E.: 'Light stemming for Arabic information retrieval': 'Arabic computational morphology' (Springer, 2007),
[74] Han, J., Kamber, M., and Pei, J.: 'Data mining: concepts and techniques' (Elsevier, 2011. 2011)
[75] Frommholz, I., al-Khateeb, H.M., Potthast, M., Ghasem, Z., Shukla, M., and Short, E.: 'On Textual Analysis and Machine Learning for Cyberstalking Detection', Datenbank-Spektrum, 2016, 16, (2),

W. Alabbas, H. M. Al-Khateeb and A. Mansour, "Arabic text classification methods: Systematic literature review of primary studies," 2016 4th IEEE International Colloquium on Information Science and Technology (CiSt), Tangier, 2016, pp. 361-367. doi: 10.1109/CIST.2016.7805072